International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Automatic Student Analysis and Placement Prediction using Advanced Machine Learning Algorithms

Kachi Anvesh, B. Satya Prasad, V. Venkata Sai Rama Laxman, B. Satya Narayana

Abstract:*The amenable statement with respective to Company Organization, Institution and students is that the company organization are taking more time to recruit which is a big challenge to them and there is no specific platform to recruit candidates on preferred qualifications. The Institutions are unable to get 100% placements among eligible students.The institutions doesn't provide proper training on minimum and preferred qualifications to the students. The candidates are unable to get specific training from college organization. The college organization should provide the training to candidates at what they are lagging behind and make the students to get stronger in preferred qualifications and all other aspects. "52% of Talent Acquisition leaders say the hardest part of recruitment is screening candidates from a large applicant pool". The time spent on screening students from a large pool often takes up the largest portion of the time. Despite College Organization are also not training students effectively based on company requirements. The analysis of student is to be done to know where the student is failing to get the placement. The company doesn't know the personalityof the student while recruiting the students. To solve this bottleneck in recruiting we created this automation tool. The main process of determining whether a candidate is qualified based on minimum qualifications like CGPA, Certifications, Projects done, Internships and respectively.*

There are two main goals of this project are:

1. *To decide whether to move the student forward to an interview or to reject them.*
2. *The college organization can give more training to the students those who got rejected by small issues like communication, programming, aptitude…*

This process is based on minimum qualifications and preferred qualifications. Both types of qualifications are more useful to the recruiters. These qualifications can include experience on projects, education, skills and knowledge, personality traits, competencies. The minimum qualifications are the mandatory qualifications that the company organizations required and preferred qualifications are not mandatory but to make the student stronger from other students. The personality and also technical knowledge can be given accurately by the faculty, mentor, H.O.D.

Based on the qualifications, personality analysis, we can shortlist the students and proper analysis is done. So the recruiters don't spend more time and college scan improve their placement percentage to 95% by improving more skills of the students, improving skills who are lagging behind (not short-listed). The colleges can predict how many students are going to be placed and who are needed to be trained more. The students can evaluate themselves about their suitable job role which make organizations easy to give job role to the students. This paper mainly concentrates on the career area prediction of computer science domain candidates.

Key words: *Machine learning, Data Science, prediction, training, testing, SVM, XG Boost, Decision tree, Logistic Regression, OneHot Encoder.*

I. INTRODUCTION

Competition is increasing day to day in today's society. Especially, more in the technical field to compete and fulfill the goals of the students in every institutions. Moreover, all the students in this society are showing interest to adopt the various technical fields. So each and every institution should workout at the initial stage of student. Every institution should evaluate the student performance constantly, identifying the interests of students, how close they are to reach their goal, to determine whether they are in the right path to reach their targets and to strengthen the weak areas to be successful. To know all these there should be a pre-evaluation before starting their career goal via student is going to attend company placement.

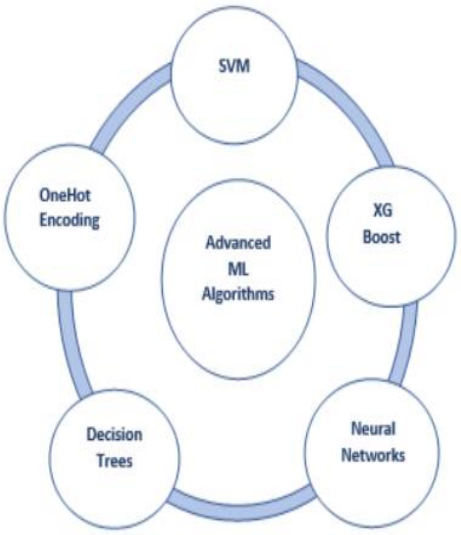
While in recruiting process, the recruiters doesn't focus on only one particular domain knowledge because there are many type of roles like software engineer, Technical support, Network Engineer, Business Analyst, Web Developer, Software Tester, Data Scientist, Database Administrator and so on. So, a recruiter analyzes and evaluates every student in all areas, interests in particular domain and then place the student in right role convenient for him. Any recruiter does not take decision that this role is best suited for him. Though there were many third party online portals which analyses the student and gives a role based on his performance. But, this is a wrong analysis because there are other parameters to be evaluated. For example, AMCAT is a platform that evaluates the student only on some parameters like aptitude, verbal skill, Reasoning and technical questions on C, sql and java. But when a student aims on role like Data Scientist, he should be evaluated on the other parameters like technical questions on Data Science, Machine Learning, python, hobbies, certifications, internships, behavior, working nature and so on.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Kachi Anvesh*, Assistant Professor, Department of IT, Vardhaman College of Engineering, JNTUH, Hyderabad, India, Email: anveshitse@gmail.com
B Satya Prasad, Student, pursuing final year, Department of IT, Vardhaman College of Engineering, JNTUH, Hyderabad, India , Email: prasadsatya952@gmail.com
Venkata Sai Rama Laxman, Student, pursuing final year, Department of IT, Vardhaman College of Engineering, JNTUH, Hyderabad, India, Email: sairamalaxman@gmail.com
B Satyaarayana, Student, pursuing final year, Department of IT, Vardhaman College of Engineering, JNTUH, Hyderabad,India, Email: satyanarayanagajala@gmail.com

So to evaluate a student we should consider totally 36 parameters and we get job roles as a result of totally of 15..

S. No.	Name of the Paper	No. of Questions	Time Duration
1.	**English Comprehensive**	18	16
2.	**Quantitative Aptitude**	16	18
3.	**Logical Ability**	14	16
4.	**Aspiring Minds Personality Inventory (AMPI)**	90	20
5.	**Optional Module**		
	Computer Science	26	25
	Electronic and Semiconductors	25	35
	Telecommunication	25	30
	Electrical Engineering	25	30
	Mechanical Engineering	30	25
	Civil Engineering	40	25

As there are many input parameters and class labels normal algorithms cannot give the best output for analysis, classification and prediction. Data Science is used for preprocessing the data and analyzing the data. Machine learning algorithms like SVM, Logistic Regression, Random forest, Decision Tree, XG Boost are used for classification and prediction of class label. In [2] it was proved that we can use machine learning for candidate selection.



Data Science is the study of extracting insights from huge data by algorithms and scientific methods. It is used to extract knowledge from organized data and unorganized data with the help of data science. One can work on large amounts of data, analysis the data and visualize the data. Machine Learning is the field that makes computers to learn without programmed. Machine Learning is mainly used to combine statistical tools with the data to predict the class label. Few examples are recommendation systems, self driven cars, automation and so on.

In [4] they explained the concepts how to predict the student performance using the machine learning algorithms. The solutions of various problems can be solved by supervised learning, unsupervised learning and semi supervised learning. If the final class label is known priorthen it comes under supervised learning. Example is classification algorithms. If the final class label is not known it comes under unsupervised learning. Example is clustering algorithms. Finally this paper mainly machine learning algorithms for classification, prediction and analyzing algorithm performance.

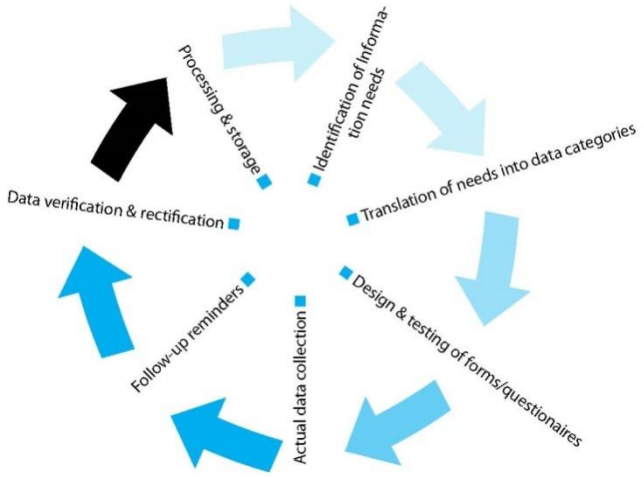
II. IMPLEMENTATION

2.1 Data Collection:

Data Collection is one of the most important tasks in machine learning projects because we are giving the input to the machine learning as data. Based on input data, the machine learns and predicts the outputs.

The Algorithm accuracy and algorithm efficiency depends on the quality, correctness, accuracy of data selected. In this prediction, many input parameters are used like student percentage in academics and working hours per day, communication skills, logical skills, coding skills, hackathon, courses, certification, workshops, interested area, interested domains, working nature and many more.

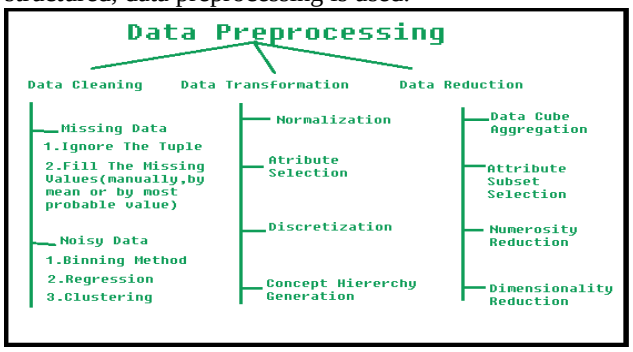
Figure 4. The process of data collection



All three parameters help us to decide student's career area. The data is collected using various ways. Some amount of data is randomly collected from different colleges. Some data is collected from company's organizations database, some data is collected from employees who are working in different organizations, and some amount of data is collected from LinkedIn. Totally, 20,000 records of data with 56 columns has been collected and organized. Data has been collected from different sources because the machine is to be trained based on the past records and predict the future by using predictive modeling.

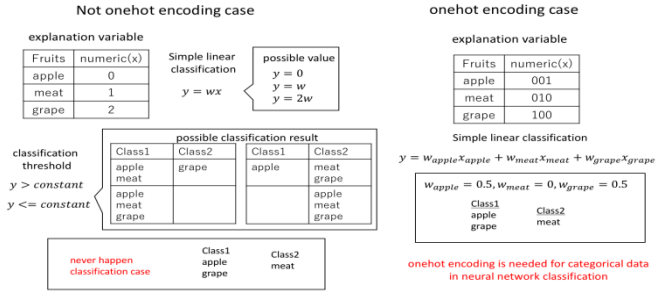
2.2 Data Preprocessing:

Collecting data is a tough task but organizing and turning the raw data into a structured format is even more difficult. In general, the data collected is having null values, anomalies and so on. To convert the unstructured data to structured, data preprocessing is used.



2.3 OneHot Encoding:

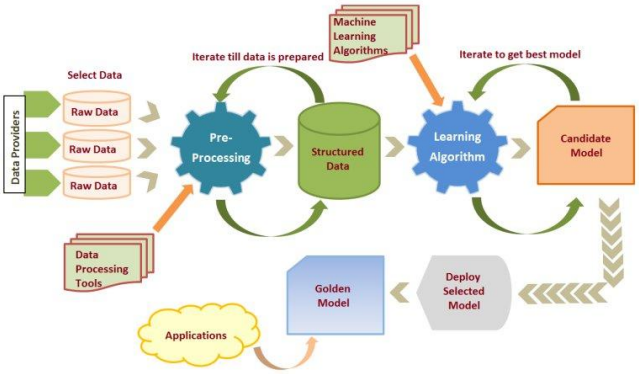
OnHot Encoding is a process that converts the categorial values in the collected data to the numerical or other ordinal values. These converted values are provided to machine learning algorithms for the prediction. This is required because many machine learning algorithms cannot be operated on label data. The input variables and output variables need to be numeric while prediction. Though random forest, decision tree handle categorical values but we are also using SVM, logistic regression cannot handle categorical values so there is a need of conversion from categorical to numerical values. The process of One Hot encoding can be explained with an example as follow if in a data there are values like good and bad, encoder assigns values like 1 and 0. That means the value of good is 1 and bad is 0 repeated for as long as appear this type of encoding is said to be integer encoding. if there are more than two values we can assign the values by for –loop or by vector method consider an example [good, average, poor,fake] it will assign the values as [0,1,2,3]. Consider another example[good, average, good, poor,good] the encoder will assign the values as [0,1,0,2,0].



2.4Machine learning algorithms:

As mentioned earlier the machine learning mechanism are of two types. They are supervised learning and unsupervised learning. Supervised learning is a learning in which we train or teach the machine using data with well labeled. After that, the machine is provided with new data so that the supervised learning algorithm analyses the data from training and produce correct result from labeled data. Supervised learning isclassifiedinto two categories they are Classification and Regression. Unsupervised learning is training the machine without any labeled data here in unsupervised algorithms the data are organized into groups based on their similarities and their patterns without any training of data. The unsupervised learning of algorithms are classifies into two categories.

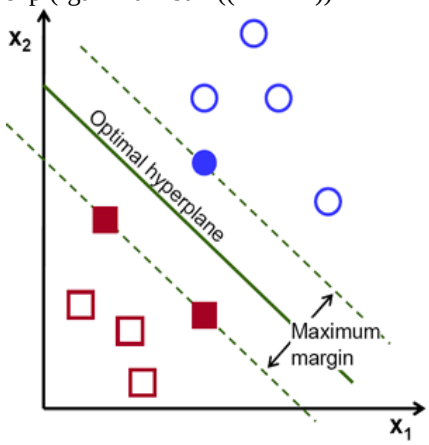
Learning analytics of mobile and ubiquitous learning environments from the perspective of human computer interaction [7,8] They are Clustering and Association.



2.4.1 SVM:

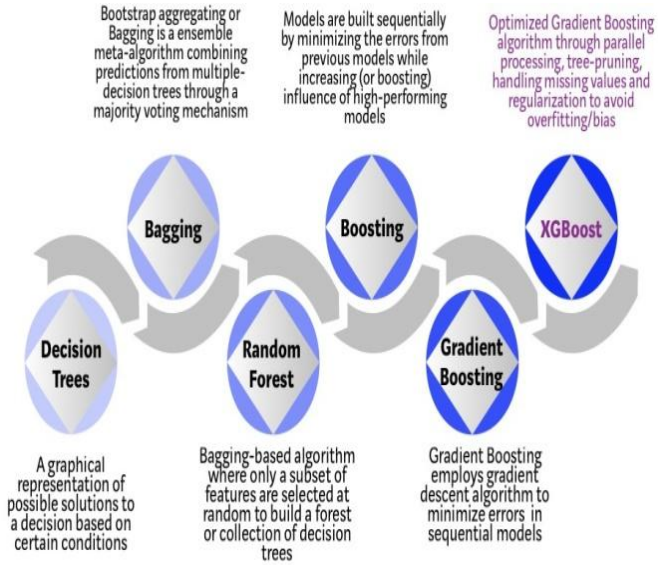
SVM means support vector machine. It is a supervised learning algorithm which is used for both classification and regression problems. It is mainly used as a classifier by separating hyperplane. From [3] we can infer that classifier can be used for the prediction purpose. In [4] we have studied different classification algorithms. In 2-D the hyperplane is represented by a line dividing two parts, each having its own class. Support vectors are the co-ordinates of individual observation. It can be used when we want to separate two classes. They are practically implemented using kernels.

The equation for dot product of an input xi and support vector xi is:
F(x) = B0 + sum(ai * (x,xi)).
Instead of using the dot-product, a polynomial kernel can be used, for example:
K(x,xi) = 1 + sum(x * xi) ^d
And not only that a more complex radio kernel is there. The general equation is:
K(x, xi) = exp (-gamma * sum((x –xi^2))



2.4.2 XG Boost:

XG Boost means eXtreme Gradient Boosting. It is the implementation of gradient boosting algorithms. It mainly focuses on performance of model and time for computation. Best features by the implementation of the algorithm are: Automatic handling of missing values.Gradient boosting is a technique where new models are made that can predict the errors or remains of previous models and then added together to make the final prediction. In training an objective function is defined. Obj=i=1 ∑l(yi,yi^(t))+ i=1 ∑ Ω(fi)

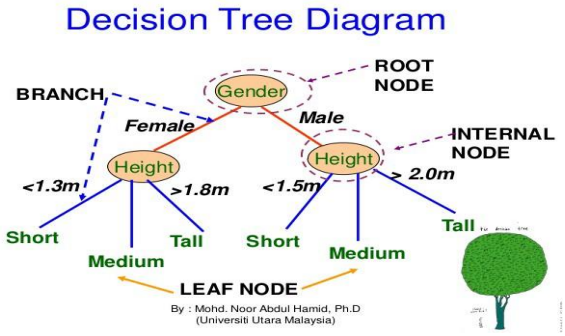


2.4.3 Decision Tree:

Decision tree is the most powerful and popular tool for classification andprediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.A tree can be “learned” by splitting the source set into subsets based on an attribute value test.This process is repeated on each derived subset in a recursive manner called recursivepartitioning. The recursion is completed when the subset at a node all has the same value of thetarget variable, or when splitting no longer adds value to the predictions. The construction ofdecision tree classifier does not require any domain knowledge or parameter setting, andtherefore is appropriate for exploratory knowledge discovery. Decision trees can handle highdimensional data. In general decision tree classifier has good accuracy. Decision tree inductionis a typical inductive approach to learn knowledge on classification.

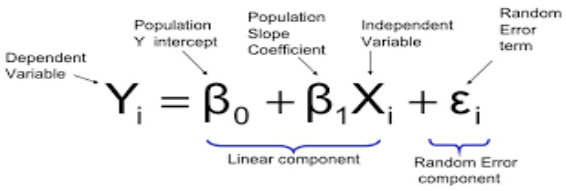
$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

$$IG(A,S) = H(S) - \sum_{t \in T} p(t) H(t)$$

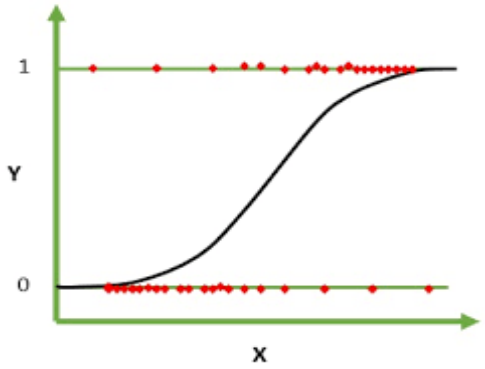


2.4.4 Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.The binary logistic regression model has extensions to more than two levels of the dependentvariable: categorical outputs with more than two values are modeled by multinomial logisticregression, and if the multiple categories are ordered, by ordinal logistic regression, for example theproportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can beused to make a classifier, for instance by choosing a cutoff value and classifying inputs withprobability greater than the cutoff as one class, below the cutoff as the other; this is a common wayto make a binary classifier.



$$\Rightarrow p(X) = \frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)}+1}$$
$$\Rightarrow p(e^{(\beta_o+\beta_1 x)}+1) = e^{(\beta_o+\beta_1 x)}$$
$$\Rightarrow p.e^{(\beta_o+\beta_1 x)} + p = e^{(\beta_o+\beta_1 x)}$$
$$\Rightarrow p = e^{(\beta_o+\beta_1 x)} - p.e^{(\beta_o+\beta_1 x)}$$
$$\Rightarrow p = e^{(\beta_o+\beta_1 x)}(1-p)$$
$$\Rightarrow \frac{p}{1-p} = e^{(\beta_o+\beta_1 x)}$$
$$\Rightarrow \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$$

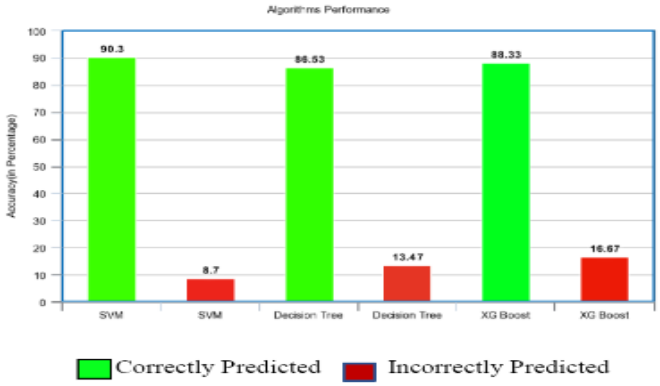


III. Training and Testing:

After processing the data(once the data is organized) the next task is to train the data and followed by testing. Once these tasks are performed we can evaluate the performance of algorithm and output is obtained. The data collected is huge, from that data 80 percent is used for training purpose and remaining 20 percent data is used for testing. The training is done to make machine to learn and giving the capability to predict the future. Whereas testing means whether is model is working properly or giving right prediction or not. This can be checked by the testing data is predefined with correct output is compared with predicted output. If there are more predefined output and predicted output are similar then the performance is more. If the accuracy is more, then continue with the model otherwise change the model.

2.6 Results:

The data is first trained and then tested with all four algorithms and out of all SVM gave more accuracy with 90.3, XG Boost with 88.33, Logistic regression with 86.3 percent accuracy. So we can conclude that SVM is the best chosen for data prediction because of having its high accuracy.



2.7 Future Scope:

A web application is developed where students cannot give inputs directly without evaluation. Once the inputs are evaluated the inputs are given ad parameter and stored in csv file. With this application we had so many advantages like

1. The students can evaluate themselves about their suitable job role.
2. The students can analyze about their strengths and weakness. They can improve the weak points and get success for their goals.
3. The company can recruit the students based on preferred qualifications and minimum qualifications.
4. The screening of students by recruiters often take less time instead of spending lots of months to give a role to the students.
5. The college institutions can achieve 90 percent of students get placed into companies.

REFERENCES

1. Ali Daud, Naif Radi Aljohani, "Predicting Student Performance using Advanced Learning Analytics", 2017 International World Wide Web Conference Committee (IW3C2).
2. Marium-E-Jannat,SaymaSultana,Munira Akther, "A Probabilistic Machine Learning Approach for Eligible Candidate Selection", International Journal of Computer Applications (0975 – 8887)Volume 144 – No.10, June 2016
3. Ms. Roshani Ade,Dr. P. R. Deshmukh, "An incremental ensemble of classifiers as a technique for prediction of student's career choice", 2014 First International Conference on Networks & Soft Computing
4. Ali Daud , Naif Radi Aljohani , "Predicting Student Performance using Advanced Learning Analytics"
5. Patricio Garcia,Analia Amandi,Silvia Schiaffino, Marcelo Campo, "Evaluating Bayesian networks' precision for detecting students' learning styles", Computers & Education ,49, pp.794- 808,2007.
6. Shabia Shabir Khan, Mushtaq Ahmed Peer, "Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013 ISSN: 2277 128X.
7. N. R. Aljohani and H. C. Davis, "Learning analytics in mobile and ubiquitous learning environments," in 11th World Conference on Mobile and Contextual Learning, 2012
8. N. R. Aljohani, H. C. Davis, and S. W. Loke, "A comparison between mobile and ubiquitous learning from the perspective of human-computer interaction," International Journal of Mobile Learning and Organization, vol. 6, no. 3/4, pp. 218- 231, 2012.

AUTHORS PROFILES

Kachi Anvesh, M Tech, have published twelve papers till now in various journals and five conferences, currently working as Assistant Professor in the department of IT at Vardhaman College of Engineering, Hyderabad Email:anvesh@vardhaman.org

Buddavarapu Satya Prasad, currently pursuing B.Tech in the field of Information Technology at vardhaman college of engineering, Hyderabad
Email: prasadsatya952@gmail.com

Venkata Sai Rama Laxman, currently pursuing B.Tech in the field of Information Technology at vardhaman college of engineering, Hyderabad Email: sairamalaxman@gmail.com

Balaga Satyanarayana, currently pursuing B.Tech in the field of Information Technology at vardhaman college of engineering, Hyderabad Email: satyanarayanagajala@gmail.com